

Application of Data mining in predicting cell phones Subscribers Behavior Employing the Contact pattern

Rahman Mansouri

Faculty of Postgraduate Studies Department of
Computer University of Najaf Abad Islamic Azad
Najaf Abad , Iran
r_mansouri@iaun.ac.ir

Mohamad Saraee

Department of Electrical and Computer Engineering
Isfahan University of Technology
Isfahan,Iran
saraee@cc.iut.ac.ir

Rasoul Amirfattahi

Department of Electrical and Computer Engineering
Isfahan University of Technology
Isfahan,Iran
fattahi@cc.iut.ac.ir

Abstract— following telecommunication services becoming competitive, client contract management in this sector has again much importance. Regarding the fact that a huge volume of telecommunication data especially details of the cell phone conversations exist, and they are practically not used, employment of data mining techniques on such data lead to exploring the hidden knowledge in them on the subscriber's behavior and lead s to predicting their behavior. Therefore, data mining is one of the most crucial methods of scientific management effective on contact with client increased profitability and client satisfaction.

In this paper, using details of the phone cell conversations during two periods (one with no rival and the second one with rival) and the employing details of conversations of the third period for identifying subscribes suffering churn, it has attempted, regarding the pattern of client to predict their churn.

Keywords: *telecom; data mining; CRM;contact pattern*

INTRODUCTION

One of the industries that has become very competitive and it new client absorption expense of which has increased as well as the expense of keeping present clients is the cell phones. Price of maintaining a present client (subscriber) is much less than that of registering a new subscriber. Certainly as we will see at present due to weakness of the second operator (provider) in external network cover and lack of service provision in most parts of the country, the number of Subscribers changing operator network this threat will become very severe.

This article is composed of two parts. The first part including section 1 to 4 covers the data mining and then the customer relationship management is described.

Then explanations will be provided on data mining procedures employed in the research and finally the employed tool is introduced.

In the second section of the paper which includes the section 5 on, the used statistical population, data collection, preprocessing and preparation of data will be explained. Then the data mining algorithm will be executed on the statistical data. And then efficiency of algorithm will be compared and obtained results will be studied and compared and finally we will conclude from our research.

1) Data Mining

What is data mining? Data mining the appropriate process for identifying patterns, rules and understandable models of data and is a stage of knowledge discovery. Data mining is an effective, advanced technique for hidden cases and is a valuable model or rule obtained from the set of many data and can perform six different jobs: classification, estimation, prediction, group dependence, clustering and description. In this research we have used the predication ability of Data mining.

Data-mining and statistics: Various sciences and technologies are used in data mining including databases, machine leaching, model identification, artificial intelligence, nervous networks, genetics and statistics. No doubt the most crucial of such sciences is the classic statistics. No data-mining is possible without statistics and statistics is the basis and foundation for most of the data mining technologies. Data mining has much in common with statistics , however, from some points differ form it, the major such differences being that in statistics a hypothesis is raised and it is either proved or rejected using statistical analysis. But in Data-mining most of the times we don not know what we are searching for.

If important covered relation existed this appear and discover. The major differ is statistical uses only numeric data but data mining uses all type of data. Then statistical is the basis of data mining but not all of it.

2) Customer Relationship Management

Customer Relationship Management (CRM) is part of automation software and related with sales and customers

and is the term used for the business practice and associated tools and infrastructure allowing business that have more than a few customers to better serve and manage the interactions with those customers the bottom line is that if you want to improve customer profitability, you almost always have to first improve the relationship that your company has with that customer. The best way to improve profitability is to improve customer loyalty and reduce churn of them.

3) Data Mining Methods

There are many methods and algorithms for Data mining. Regarding the fact in this research we will predict Cell phone subscribers Behavior employing the detail of Contact pattern. Using the Decision Tree is better. We use four algorithm CART, CHAID, QUEST, C5.0 when all of the Decision Tree algorithms.

Some of the characteristics of them are follows.

1.3: CART algorithm: The algorithm is based on Classification and Regression Trees by Breiman et al (1984). A Classification and Regression Tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The method uses recursive partitioning to split the training records into segments by minimizing the impurity at each step. The Classification and Regression Tree node generates a decision tree that allows you to predict or classify future observations. Target and predictor fields can be range or categorical and all splits are binary (only two subgroups).

2.3: CHAID algorithm: The CHAID technique was created by Gordon V. Kass in 1980. CHAID is a technique of decision tree or regression tree. CHAID is the best tool used to discover the relationship between variables. The CHAID node generates decision trees using chi-square statistics to identify optimal splits. Unlike the C&RT and QUEST nodes, CHAID can generate no binary trees, meaning that some splits have more than two branches. Target and predictor fields can be range or categorical.

3.3: QUEST algorithm: QUEST is a binary-split decision tree algorithm for classification and data mining developed by Wei-Yin Loh (University of Wisconsin-Madison) and Yu-Shan Shih (National Chung Cheng University, Taiwan). QUEST stands for Quick, Unbiased and Efficient Statistical Tree. The QUEST node provides a binary classification method for building decision trees, designed to reduce the processing time required for large C&RT analyses while also reducing the tendency found in classification tree methods to favor predictors that allow more splits. Predictor fields can be numeric ranges, but the target field must be categorical. All splits are binary.

4.3: C5.0 algorithm: Quinlan went on to create C5.0 and See5 (C5.0 for Unix/Linux, See5 for Windows) which he markets commercially. C5.0 offers a number of improvements on C4.5. Some of these are
Speed - C5.0 is significantly faster than C4.5 (several orders of magnitude)

Memory usage - C5.0 is more memory efficient than C4.5

Smaller decision trees - C5.0 gets similar results to C4.5 with considerably smaller decision trees.

The C5.0 node builds either a decision tree or a rule set. The model works by splitting the sample based on the field that provides the maximum information gain at each level. The target field must be categorical. Multiple splits into more than two subgroups are allowed.

4) Instrument used

To implement and apply the above said algorithm we used the Clementine tool version 11.1. The reason for using this tool was comprehensively of models, user friendly media, ease of data analysis, low expense and

5) Statistical Population

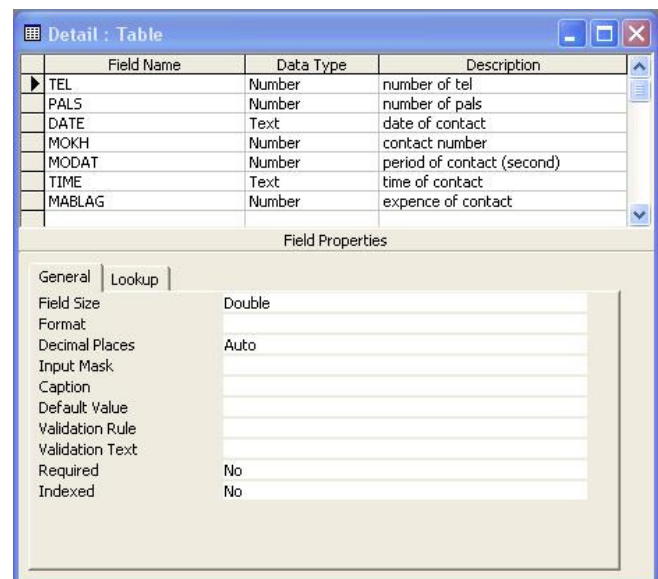
permanent cell phones and credit cell phones on one of the telecommunication center \s of the Chaharmahal-and-Bakhtiyari including 928 cell phones (1st permanent) and 1727 credit 1st cell phones.

6) Data collection

Data was gathered relevant to details of the telephone conversations during three temporal periods. The first period concerns details of conversations of February to mid April which represent only cell phone services of the telecommunication company. The second temporal period includes mid-April to the end of May when second operator of the cell phone has become active in this place and the second periods extend from late May to July 5th.

7) Preprocessing and preparation of data

Collected data includes 2655 cell phones in the 1st and 2nd periods includes 487865 information (data) records comprising details of their conversations in such temporal periods and has a chart structure as the following form.



Field Name	Data Type	Description
TEL	Number	number of tel
PALS	Number	number of pals
DATE	Text	date of contact
MOKH	Number	contact number
MODAT	Number	period of contact (second)
TIME	Text	time of contact
MABLAG	Number	expense of contact

Field Properties	
General	
Field Size	Double
Format	
Decimal Places	Auto
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	No
Indexed	No

Table1: Structure of file containing details of cell phone conversations

Regarding the fact that churn prediction of cell phones is possible using contact patterns, a field has been added to the conversations and is speaker addressed according to the type of speaker. Type of speaker is identified through comparison with this same file and files of the fixed telephones of the telecommunication center, inside and intercity communication or contact with another operator. In the

subsequent stage, summing up of such data is done based on number of cell phones and employing cumulative functions on the price and period fields and period and number of contacts based on urban, inter city, and the second operator. Then, computations were done in ratio with price, period, and number of contacts in the second period compared with the first period and was recorded in the file as new fields (based on city, intercity and operator data). Also, ratio of total sum of communications in the second period to the first one was computed and registered in a new field. A new field entitled churn was added to the file which is our prediction field. Sum of those cell phones with no communication in the third period was assumed to be 1 and sum of those with communication was assumed to be 0. Finally information regarding number of conversations, price of conversations and period of city and inter city conversations and 2nd operators of the first and second periods was normalized. It is worth mentioning that in this period no cell phone has been interrupted due to debts and cell phones with no conversation have probably changed their operators.

Structure of the obtained chart is as follows:

Field Name	Data Type	Description
TEL	Number	mobile number
NESKOL	Number	ratio with total price of contacts in the second period compared with the first period
NES_I	Number	ratio with prices of contacts with the second operator in the second period compared with the first period
NES_S	Number	ratio with price of urban contacts in the second period compared with the first period
NES_B	Number	ratio with price of inter-city contacts in the second period compared with the first period
CHURN	Number	switch provider(operator)
NOE_TEL	Text	type of mobile (permanent , credit)
n_mod_b_I	Number	period of time of inter-city contacts in first priod
n_mod_b_Y	Number	period of time of inter-city contacts in second priod
n_mod_s_I	Number	period of time of urban contacts in first priod
n_mod_s_Y	Number	period of time of urban contacts in second priod
n_mod_I_I	Number	period of time of contacts with second operator in first priod
n_mod_I_Y	Number	period of time of contacts with second operator in second period
n_no_b_I	Number	number of inter-city contacts in first period
n_no_b_Y	Number	number of inter-city contacts in second period
n_no_s_I	Number	number of urban contacts in first period
n_no_s_Y	Number	number of urban contacts in second period
n_no_I_I	Number	number of contacts with second operator in first period
n_no_I_Y	Number	number of contacts with second operator in second period
n_mab_b_I	Number	price of inter-city contacts in the first period
n_mab_b_Y	Number	price of inter-city contacts in the second period
n_mab_s_I	Number	price of urban contacts in the first period
n_mab_s_Y	Number	price of urban contacts in the second period
n_mab_I_I	Number	prices of contacts with the second operator in the first period
n_mab_I_Y	Number	prices of contacts with the second operator in the second period

Table 2: Structure of the file of details of the cell phone conversations following preparations and preprocessing.

8) Execution of data mining algorithm

1.8: CART algorithm: in this algorithm 80% of data are used as training data sets and 20% of the data are employed as test data. The Gini coefficient has been used. Also the tree has pruned from the fifth level.

Time of executing algorithm was 3 seconds and in 98% of cases algorithm arrived to a correct answer.

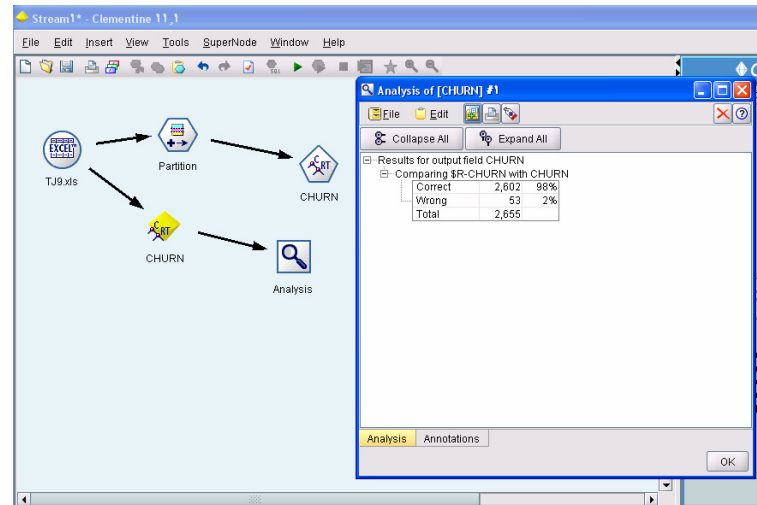


Fig 1: Execution of CART algorithm using the Clementine tool.

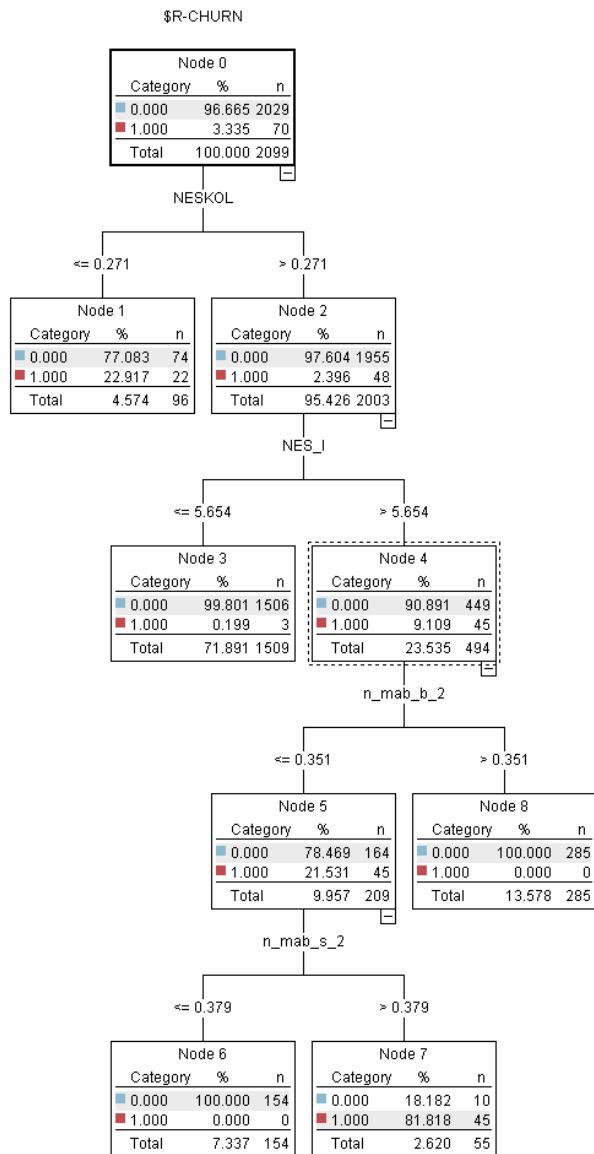


Fig2: the tree resulted form execution of CART algorithm.

Results:

Second level: In case of the telecommunication ratios, describing too much in the second period (less than 27%), probability of churn is high (about 23%). In other words in many of the common cases it will gradually suffer churn and its contact is not interrupted suddenly and at once. Probably in this case concurrently the two lines of the first cell phone and the second operator has been used however it most contacts is with the second operator line (Ratio of less than 27%)and use of the first cell phone line decreased and approaches zero.

The third level: Those subscribers whose contacts with subscribers of the second operator show sufficient increase, probably suffer more churn (9.17).

The fourth level: the more inter-city contacts, the less probability of churn. Probably for those subscribers with high inter-city communication the network cover is more crucial than expenses.

Level five: subscribers with more city communication have more churns as well.

2.8: CHAID algorithm: in this algorithm from 80% of data was used as training ones and 20% were employed as test ones and the α factor equal to 0.05 was used for split and merge and also the tree was pruned at the fifth level.

Time period for executing the algorithm was 2seconds and 97.59% of the algorithm cases arrival to a correct answer.

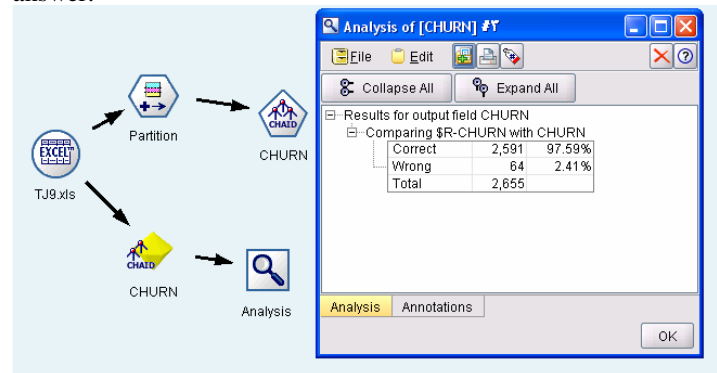


Fig 3: Execution of the CHAID algorithm using the Clementine tool.

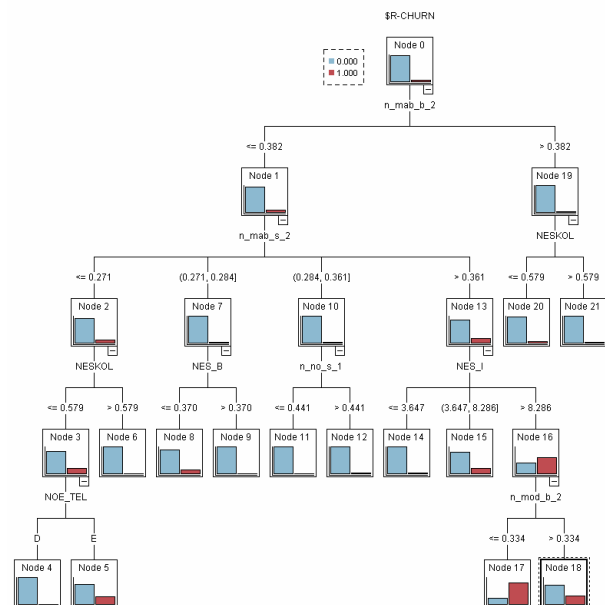


Fig 4: decision tree resulting from execution of CHAID algorithm.

Results: Results obtained from this tree are similar to those from the CART algorithm however the role of inter-city and contacts with the second operator is more prominent. Meanwhile regarding the nodes 5 and 6 it can be concluded that probability of churn in credit cell phones is more than permanent cell phones.

3.8: QUEST algorithm: in this algorithm 80% of data were used as training data and 20% were employed as test ones and the α factor equal to 0.05 was used to split data and also the tree was pruned at level five. Time for executing the algorithm is 1 second and in 96.42% of cases algorithm arrival to correct answers.

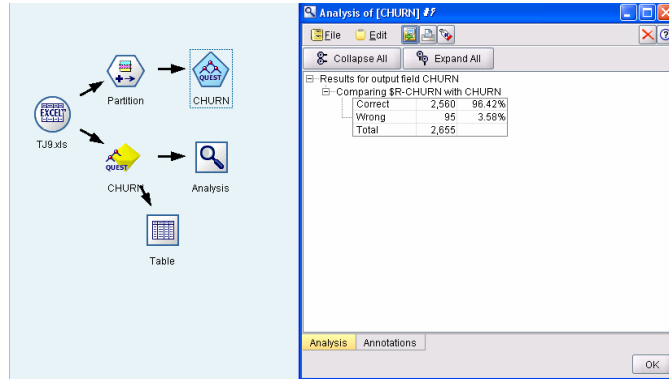


Fig 5: Execution of the QUEST algorithm using the Clementine tool.

\$R-CHURN		
Node 0		
Category	%	n
0.000	96.665	2029
1.000	3.335	70
Total	100.000	2099

Fig 6: The decision tree resulting form execution of QUEST algorithm.

Results: This algorithm predicts all telephones with no churn that regarding the fact that the number of 95 telephones from the total 2655 telephones have suffered churn and that the ratio has been computed for 80% the obtained percent is logical.

4.8. The C5.0 algorithm: in this algorithm 80% of data were used as the set of training data and 20% were employed as test ones. The algorithm execution period was 1 second and in 98.12% of cases the algorithm arrival to correct answer.

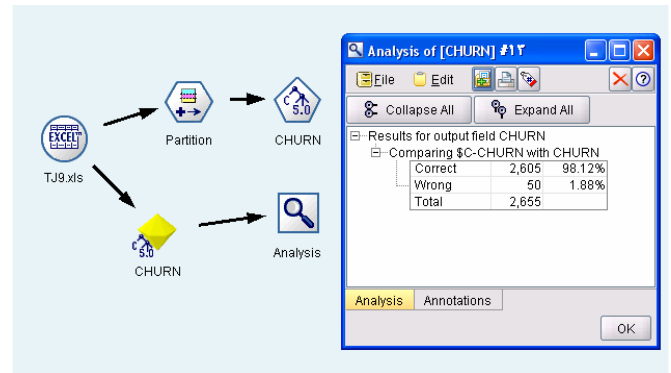


Fig7: execution of C5.0 algorithm using the Clementine tool.

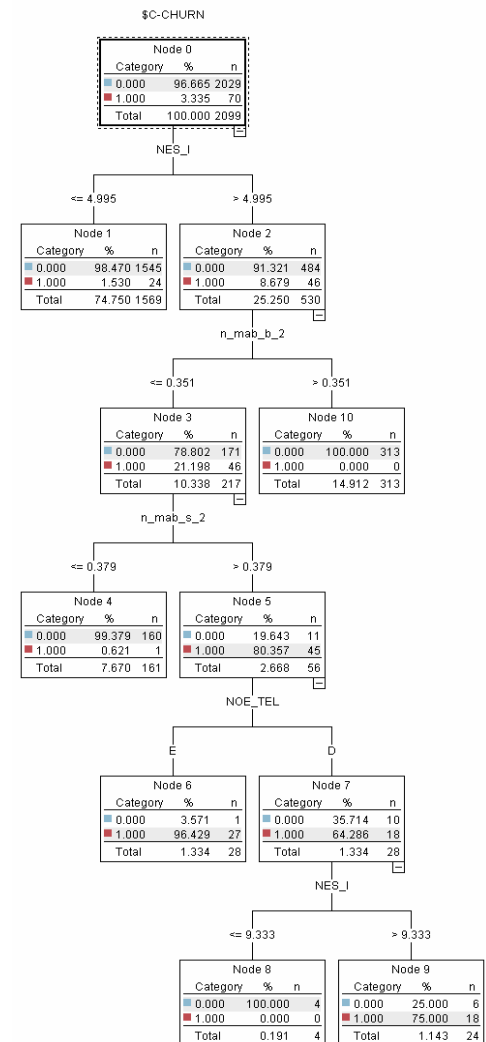


Fig 8: The decision tree resulting form execution of C5.0 Algorithm.

Results:

Second Level: Those Subscribers whose communications with the second operator subscribers had significant increase have more probability of Churn (8.7%).

The third level: The more the inter-city communications, the less probability of churn. It is probable that for a subscriber of more inter-city contacts, the network cover is more crucial than expense.,

The fourth level: subscribers with more city communication have more churns as well..

The fifth level; the credit first cell phone subscriber compared with permanent ones have more probability of churn.

The sixth level: Subscribers whose contacts with the second operator has increased have more chance of churn. These results except those of the fifth level were obtained in the CART algorithm as well and the fifth level result had been obtained in the CHAD algorithm as well.

Therefore these algorithm in addition to ease enjoys universality of the CART and CHAID algorithms as well.

9) Comparison

Comparison of the four above said algorithms based on precision, speed and complexity. (numbers of the tree's nodes) are compared in the following table.

Algorithm	Precision	Response time (second)	Complexity
CART	98%	3	Medium(9nodes)
CHAID	97.59%	2	High(21 nodes)
QUEST	96.42%	1	Low(1 node)
C5.0	98.12%	1	Medium(11node)

Table3: Comparison of four algorithms of the decision tree.

Regarding the above issues, the best algorithm is the C5.0 algorithm.

10) Conclusion and Suggestions

In this paper one of the major problems of the telecommunication Co. namely loss of subscribers was studied and procedure of data mining and appropriate algorithm for specifying at risk subscribers regarding their contact patterns was introduced. No. doubt this has been done at a time section and it is proposed , regarding dynamics and continuity of telephone conversations that this would be done continuously and the algorithms and fields effective on prediction be amended.

It is proposed that this research is reputed using other characteristics of the cell phones such as particulars and social status of owners, and technical specifications of contact i.e. ASR factor as well. It seems that the percent of telephones will churn is low (3.58%) however, it should be noted that this figure is for one and a half months and it might be continued in the next month and secondly through development of the second operator network and its overall cover the probability that a new arrange of cell phone subscribers become willing to change their operator is high and although active subscriber of the telecommunication Co are not at risk however regarding existence of the second

operator, the communication Co. has lost many potential subscribers. Comparing sale for the first permanent and credit cell phones of temporal sections that the second operator has not been active periods when the second operator has become active in the region, it is clearly evident. Therefore, the telecommunication Co. should through study of the contact pattern of the active subscribers prepare proposal for at risk subscribers and prepare various proposal for different social spectra of its potential subscribers.

REFERENCES:

- [1] Building Data Mining Applications for CRM by Alex Berson, Kurt Thearling, and Stephen J. Smith (Kindle Edition - Jan 4, 2002)
- [2] Data Warehousing and Data Mining for Telecommunications by Rob Mattison(Artech House, Inc. Norwood, MA, USA)
- [3] Data mining and surveillance in the post-9.11 environment Oscar H. Gandy,.,Herbert I. Schiller Barcelona, July, 2002
- [4] An SVM based Churn Detector in Prepaid Mobile Telephony Cédric Archaux , Hicham Laanaya , Arnaud Martin , Ali Khenchaf (2004 IEEE)
- [5] Mining Optimal Actions For Profitable CRM Charles X.ling, Qiang Yang,Jie Cheng (2002 IEEE)
- [6] A Review of Data Mining Tools In Customer Relationship Management Jayanthi Ranjan, Vishal Bhatnagar(Journal of Knowledge Management Practice, Vol. 9, No. 1, March 2008)
- [7] Research on Application of Data Mining Technology in CRM Ying Gao, Dezhen Feng(The Sixth Wuhan International Conference on E-Business e-Business Track)
- [8] Critical Success Factors For Implementing CRM Using Data Mining Jayanthi Ranjan, Vishal Bhatnagar(Journal of Knowledge Management Practice, Vol. 9, No. 3, September 2008)
- [9] Zhao Hongbo, Relationship Management of Telecom Enterprise's Customers, Apr. 2003.
- [10] Research on Applying Data Mining to Telecom CRM Peng Liu, Naijun Wu, Chuanchang Huang, Bingrong Wei, Libo Wang, Zhen'an He (International Forum of Information System Frontiers - 2006 Xian International Symposium)
- [11] Shu Huiying, Qi Jiaying, Lifespan Management of Telecom Customers, Aug. 2004.
- [12] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007